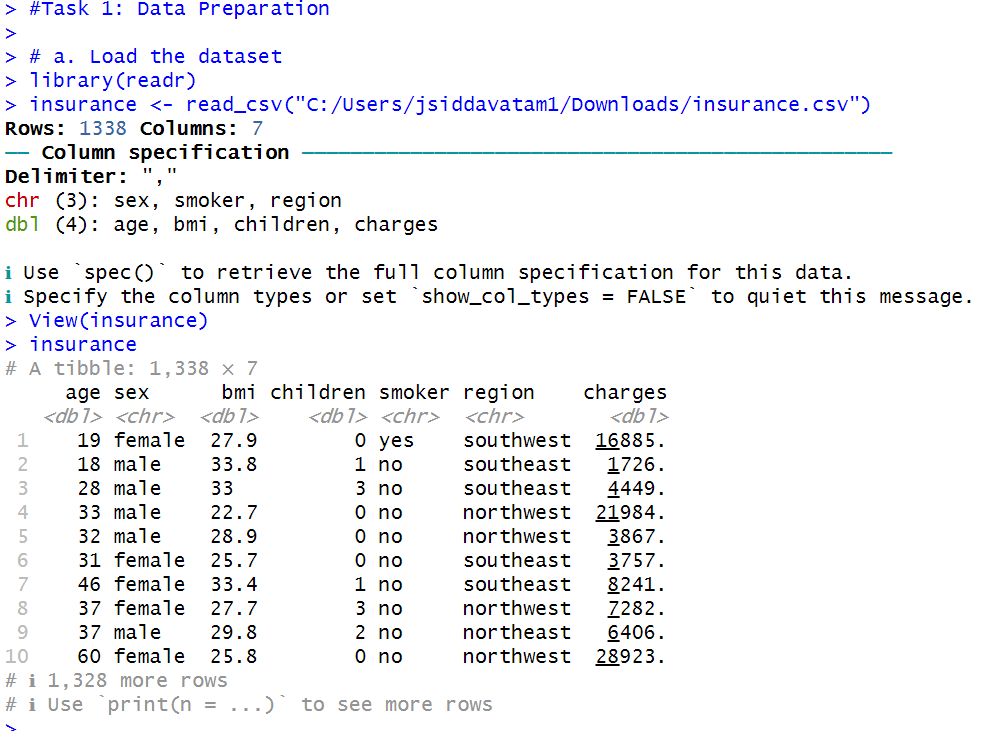
**PREDICTIVE MODELING**

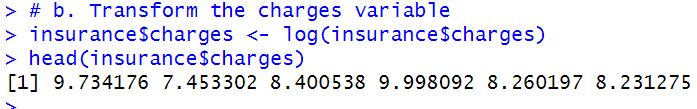
**Final Course Project**

**Name:** Ujwal Chepyala

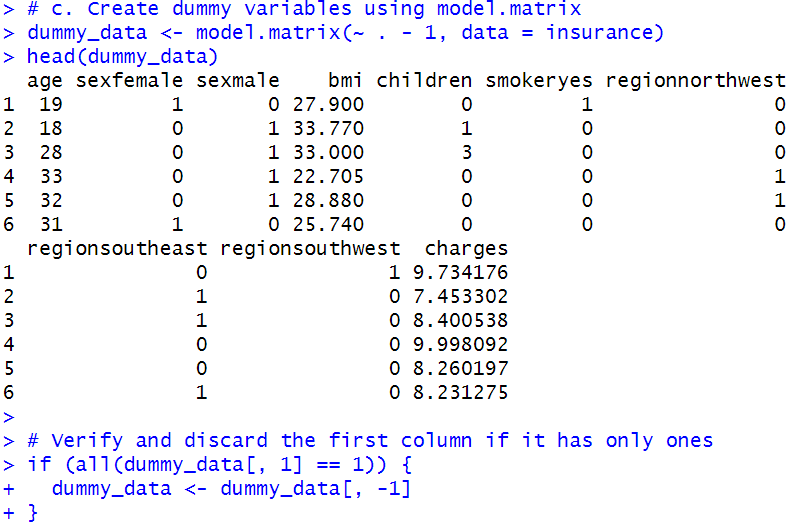
1. Data Preparation.
2. Load the dataset into memory.



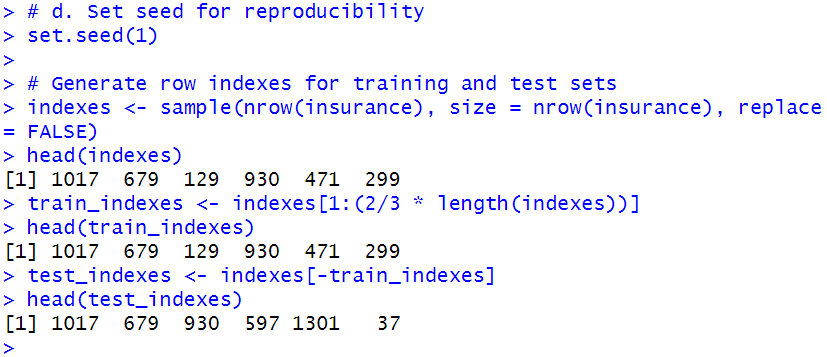
1. In the data frame, transform the variable **charges** by setting **insurance$charges = log(insurance$charges).**



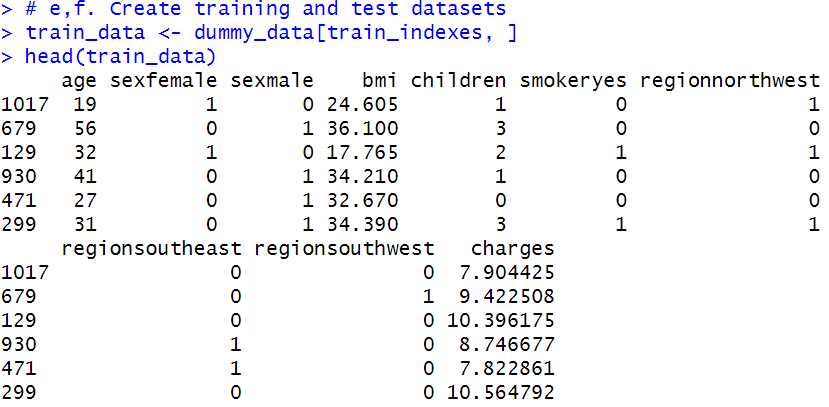
1. Using the data set from 1.b, use the **model.matrix()** function to create another data set that uses dummy variables in place of categorical variables. Verify that the first column only has ones(1) as values, and then discard the column only after verifying it has only ones as values.



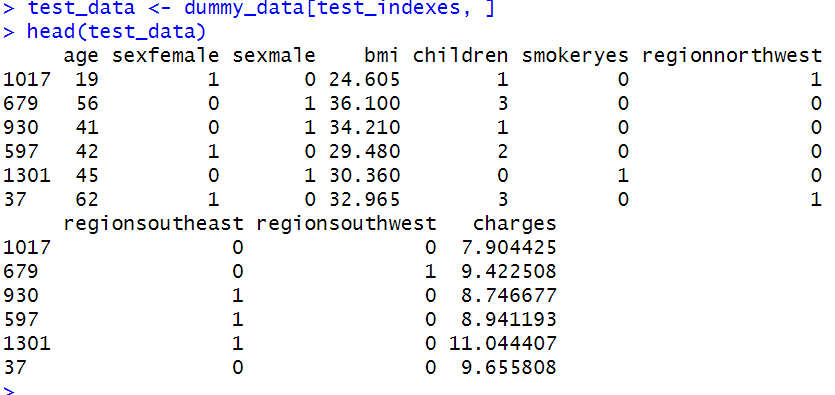
1. Use the **sample()** function with set.seed equal to 1 to generate row indexes for your training and tests sets, with 2/3 of the row indexes for your training set and 1/3 for your test set.



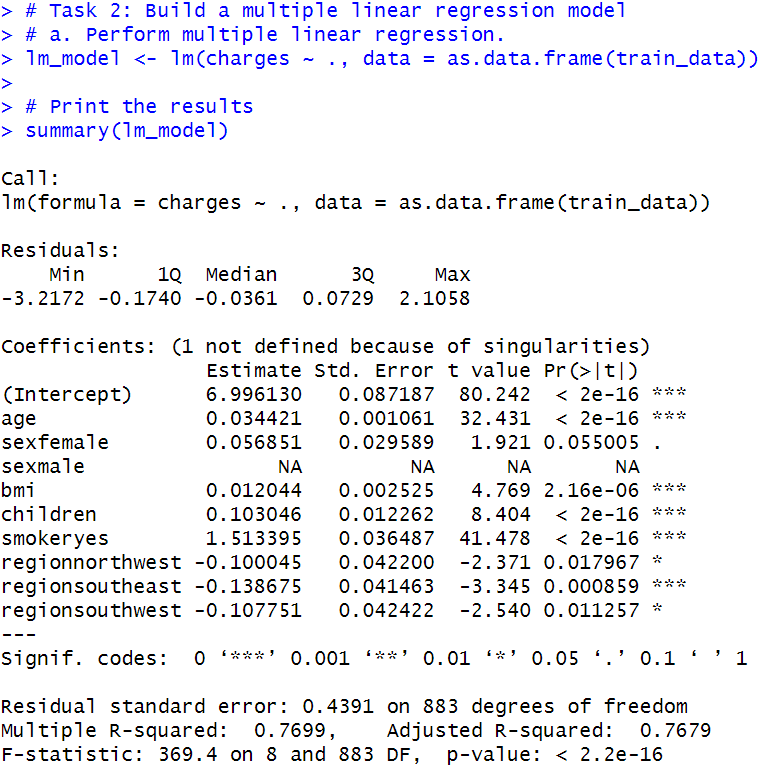
1. Create a training and test data set from the data set created in 1.b using the training and test row indexes created in 1.d.



1. Create a training and test data set from data set created in 1.c using the training and test row indexes created in 1.d.



1. Build a multiple linear regression model.
2. Perform multiple linear regression with **charges** as the response and the predictors are **age, sex, bmi, children, smoker,** and **region.** Print out the results using the summary() function.



1. Is there a relationship between the predictors and the response?

Yes, there is a relationship between the predictors and the response. The results from the multiple linear regression model summary indicate several important points:

Statistical Significance: The coefficients for variables such as age, bmi, children, smokeryes, regionnorthwest, regionsoutheast, and regionsouthwest are statistically significant. This is indicated by the low p-values (typically < 0.05), denoted by asterisks in the "Pr(>|t|)" column.

Direction of Relationship: The sign of the coefficients indicates the direction of the relationship. For example, a positive coefficient for age suggests that as age increases, charges tend to increase. Similarly, a positive coefficient for smokeryes suggests that being a smoker is associated with higher charges.

Adjusted R-squared: The adjusted R-squared value is a measure of how well the model explains the variability in the response variable. In this case, the adjusted R-squared is approximately 0.7679, suggesting that about 76.79% of the variability in charges is explained by the model.

F-statistic: The F-statistic tests the overall significance of the model. The very low p-value (p-value: < 2.2e-16) indicates that the overall model is statistically significant.

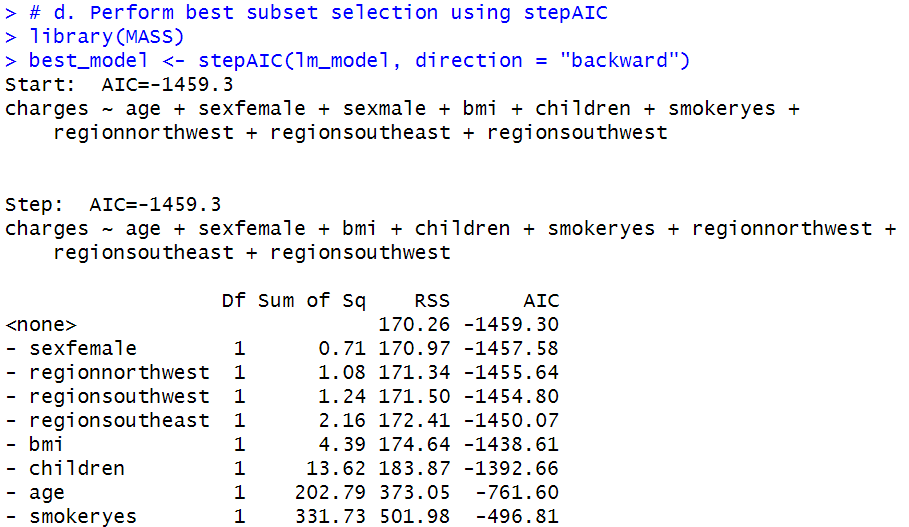
In summary, the results suggest that the predictors (age, sex, bmi, children, smoker, and region) collectively have a statistically significant relationship with the response variable charges. The model explains a significant portion of the variability in charges, as indicated by the adjusted R-squared value.

1. Does **sex** have a statistically significant relationship to the response?

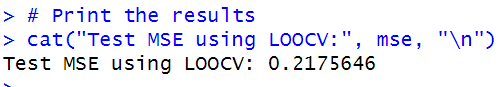
The 'sexfemale' coefficient has an associated p-value of 0.055005. This p-value is slightly above the conventional significance level of 0.05. Therefore, the relationship between the 'sex' variable (specifically 'sexfemale') and the 'charges' response is not statistically significant at the 0.05 significance level.

In practical terms, this suggests that there is some evidence (p-value of 0.055) that being female may be associated with higher charges, but this evidence is not strong enough to reach conventional levels of statistical significance. The relationship should be interpreted cautiously, and further investigation or consideration may be needed.

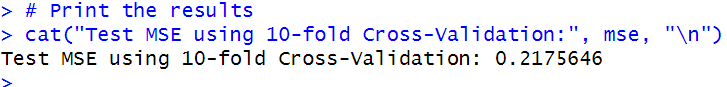
1. Perform best subset selection using the stepAIC() function from the MASS library, choose best model based on AIC. For the “direction” parameter in the stepAIC() method, set direction = “backward”.



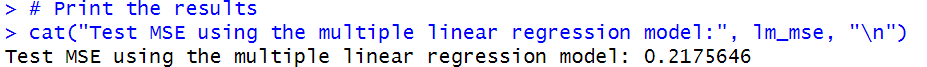
1. Compute the test error of the best model in #2d based on AIC using LOOCV using trainControl() and train() from the caret library. Report the MSE by squaring the reported RMSE.



1. Calculate the test error of the best model in #2d based on AIC using 10-flod Cross-Validation. Use train and trainControl from the caret library. Refer to model selected in #2d based on AIC . Report the MSE.



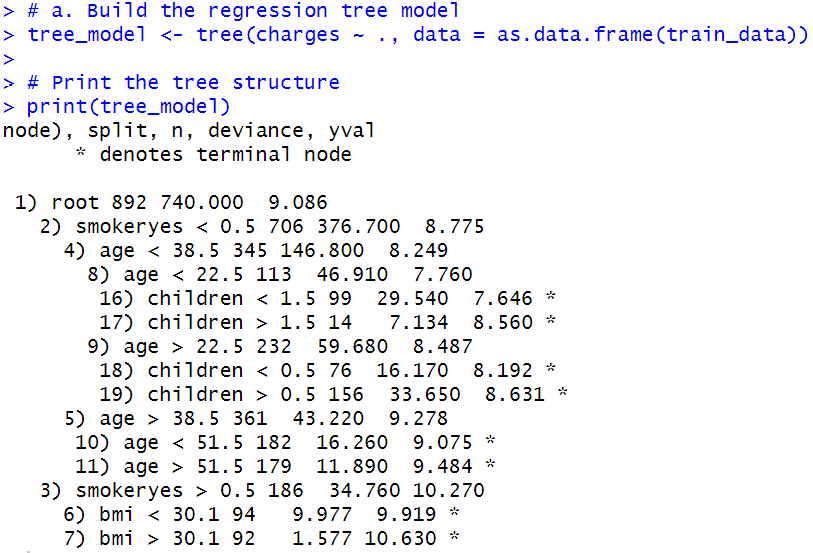
1. Calculate and report the test MSE using the best model from 2.d and test data set created in step 1.e.



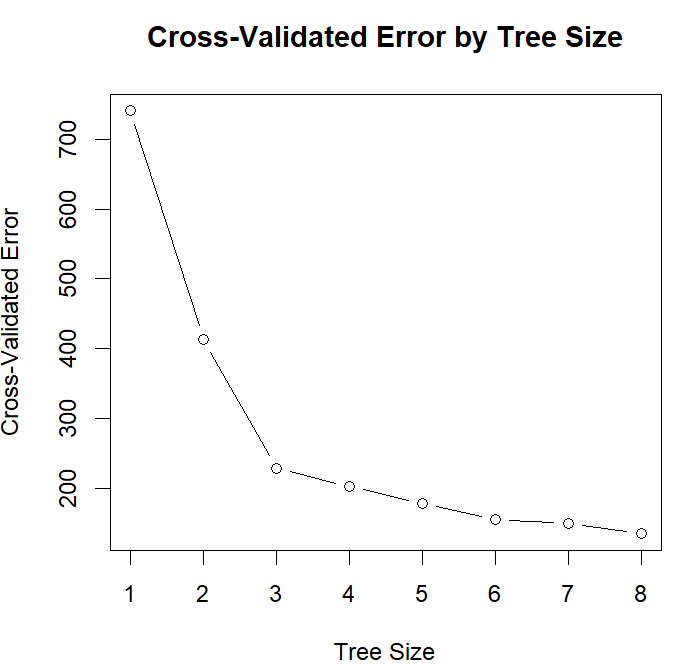
1. Compare the test MSE calculated in step 2.f using 10-fold cross-validation with the test MSE calculated in step 2.g. How similar are they?

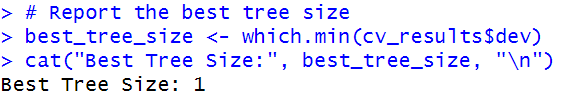
The test MSE values you provided from steps 2.f (10-fold Cross-Validation) and 2.g (LOOCV) are both 0.2175646. In this case, the two values are identical, indicating that the performance of the multiple linear regression model is consistent between 10-fold Cross-Validation and LOOCV. The similarity in the test MSE values suggests that the model's performance is stable and not significantly affected by the choice of cross-validation method in this specific instance.

1. Build a regression tree model.
2. Build a regression tree model using function tree(), where **charges** is the response and the predictors are **age, sex, bmi, children, smoker,** and **region.**



1. Find the optimal tree by using cross-validation and display the results in a graphic. Report the best size.





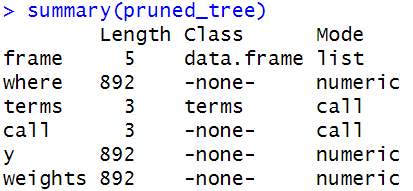
1. Justify the number you picked for the optimal tree with regard to the principle of variance-bias trade-off.

Bias: A small tree tends to have high bias because it may oversimplify the underlying relationships in the data, leading to systematic errors in predictions.

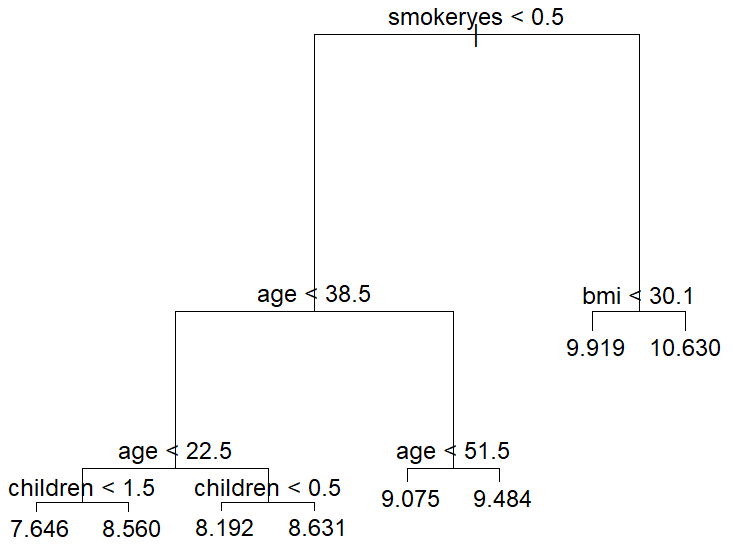
Variance: However, a small tree typically has low variance because it is less sensitive to noise in the training data.

The goal in the variance-bias trade-off is to find a balance that minimizes the overall error.

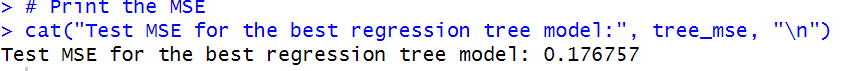
1. Prune the tree using the optimal size found in 3.b.



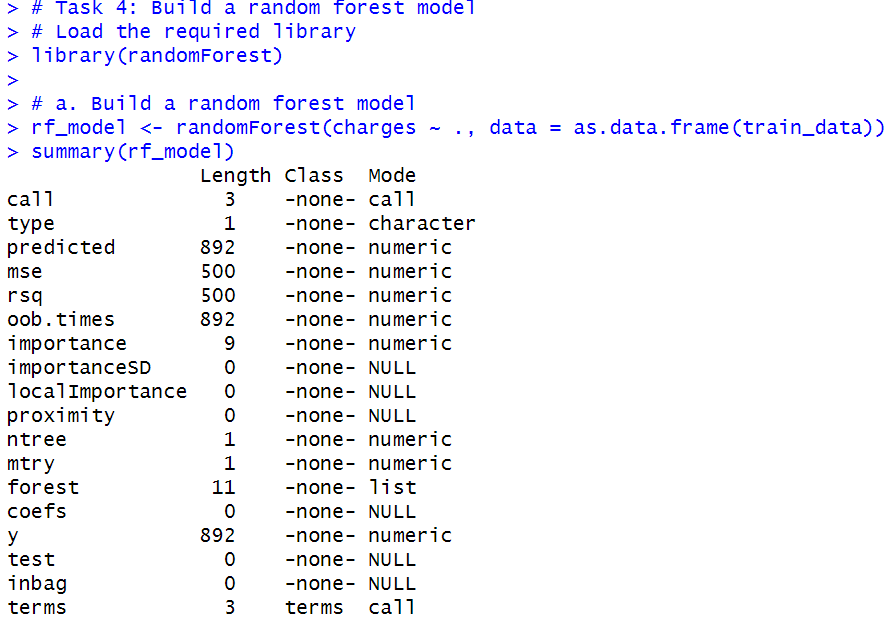
1. Plot the best tree model and give labels.



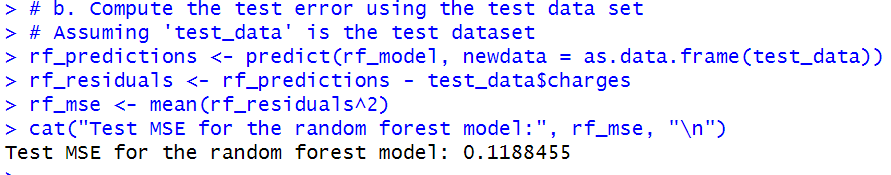
1. Calculate the test MSE for the best model.



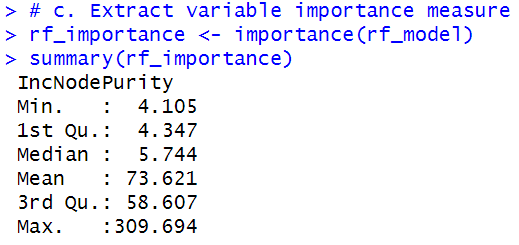
1. Build a random forest model.
2. Build a random forest model using function randomForest(), where **charges** is the response and the predictors are **age, sex, bmi, children, smoker,** and **region.**



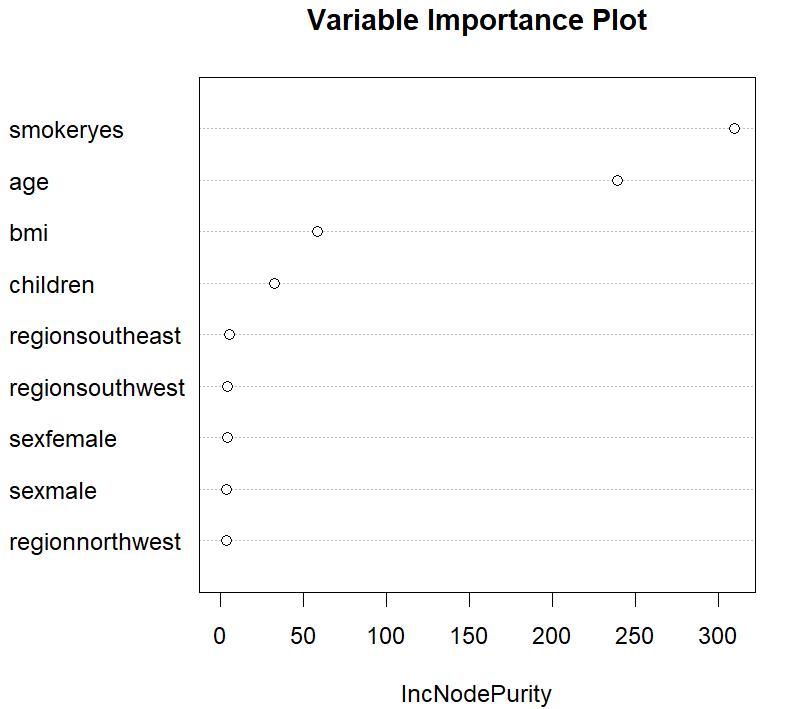
1. Compute the test error using the test data set.

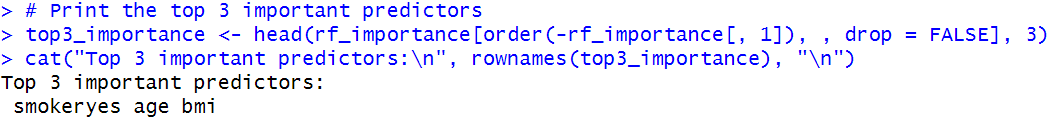


1. Extract variable importance measure using the importance() function.

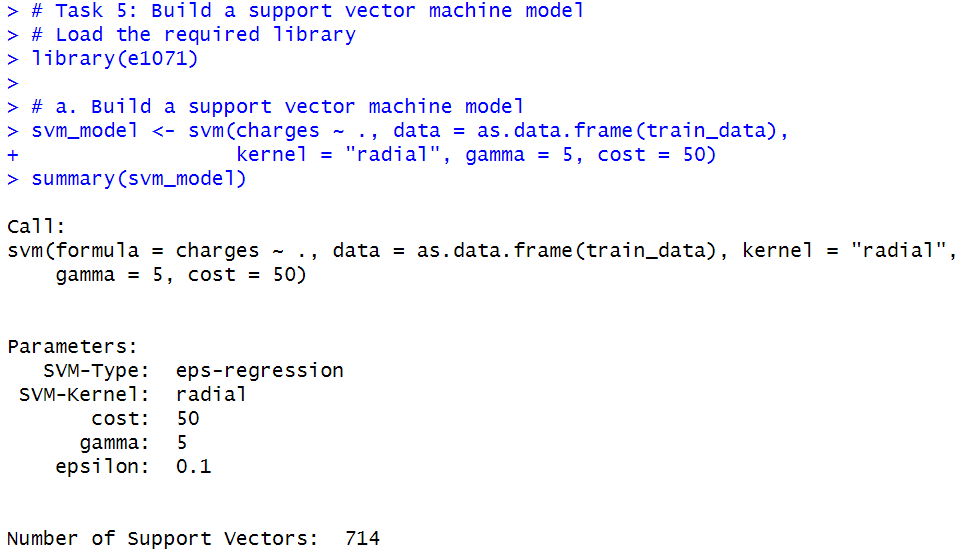


1. Plot the variable importance using the function, varImpPlot(). Which are the top 3 important predictors in this model?

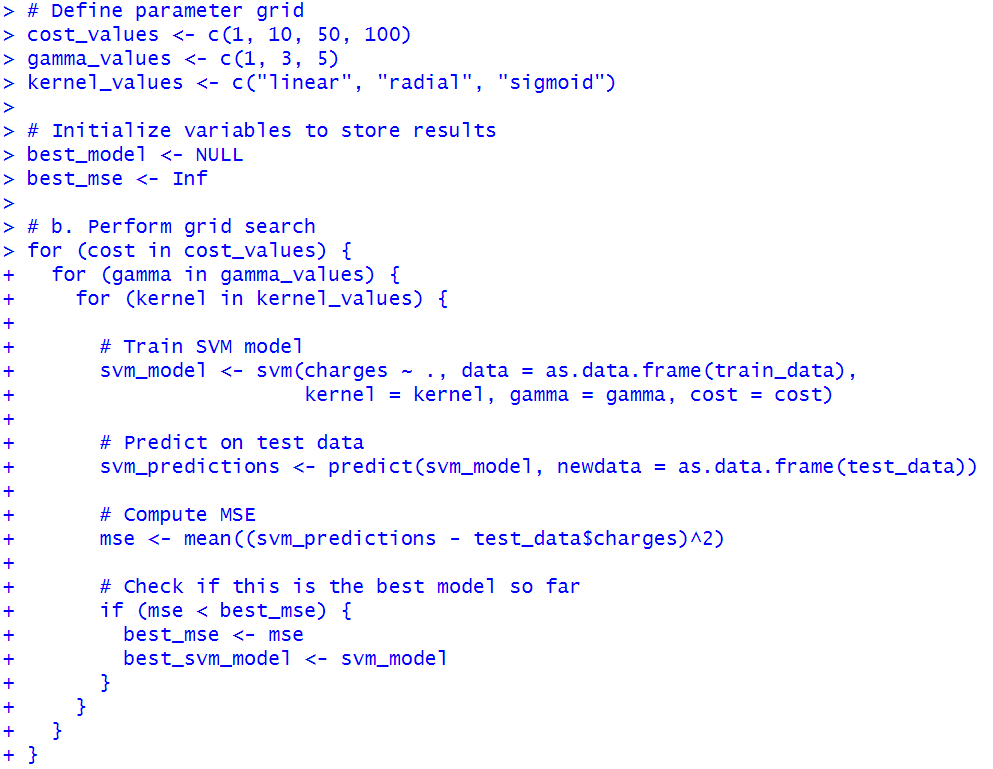




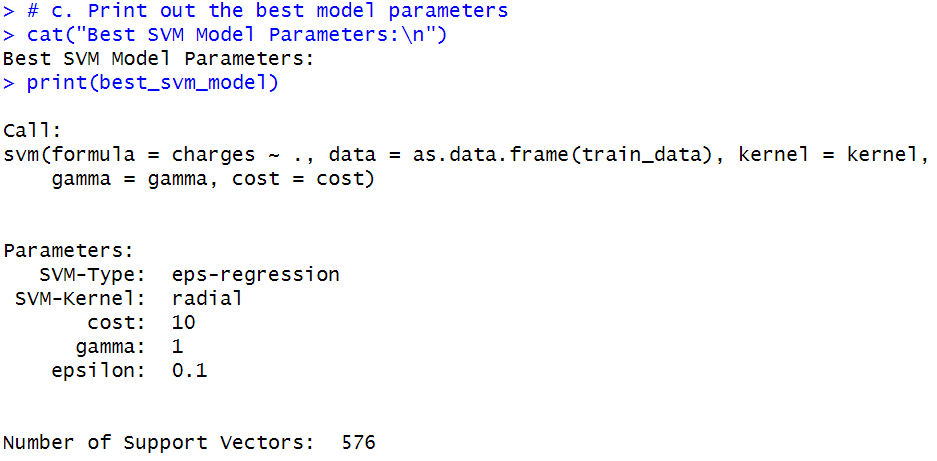
1. Build a support vector machine model.
2. The response is **charges**  and the predictors are **age, sex, bmi, children, smoker,** and **region.**



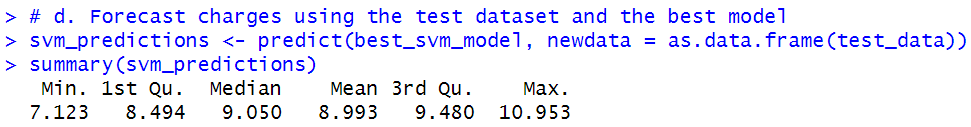
1. Perform a grid search to find the best model with potential cost: 1, 10, 50, 100 and potential gamma: 1, 3, 5 and potential kernel: “linear”, “radial” and “sigmoid”. And use the training set created in step 1e.



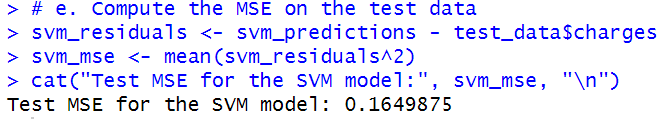
1. Print out the model results. What are the best model parameters?



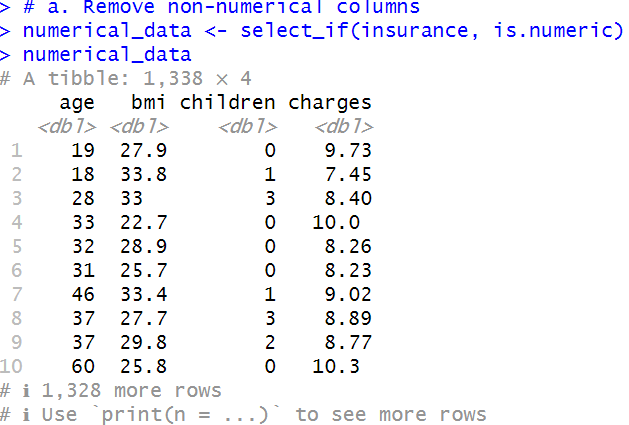
1. Forecast **charges** using the test data set and the best model found in 5c.



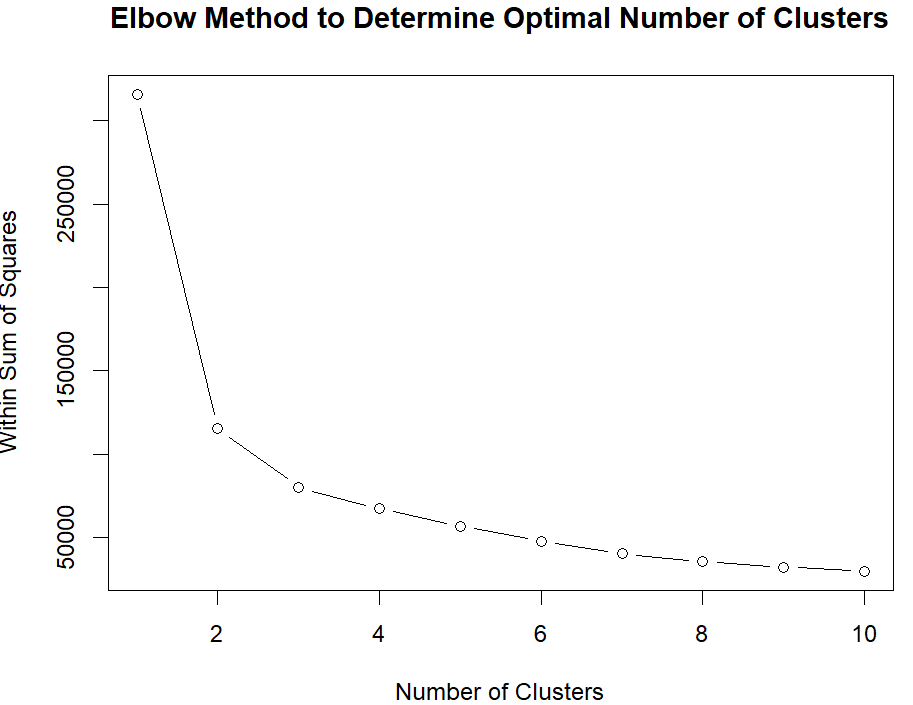
1. Compute the MSE on the test data.



1. Perform the k-means cluster analysis.
2. Remove the **sex, smoker,** and **region,** since they are not numerical values.



1. Determine the optimal number of clusters. Justify your answer.



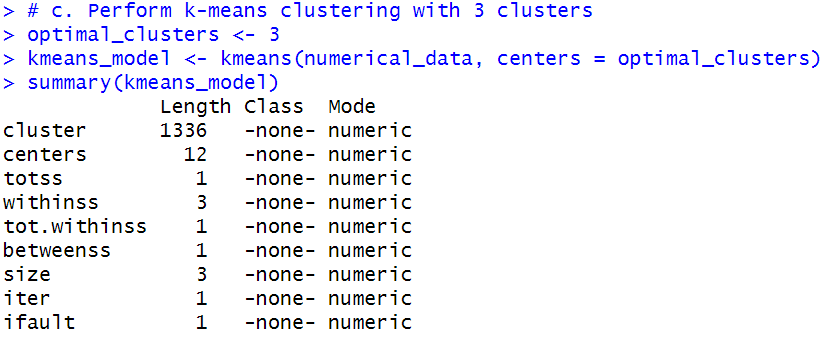
Significant Reduction Initially: Initially, as the number of clusters increases from one to three, you observe a significant reduction in the within-cluster sum of squares (WSS). Each additional cluster helps to better capture the variance in the data.

Diminishing Returns After Three Clusters: After reaching three clusters, the rate of decrease in WSS slows down. Adding more clusters does not lead to a proportionally significant reduction in WSS.

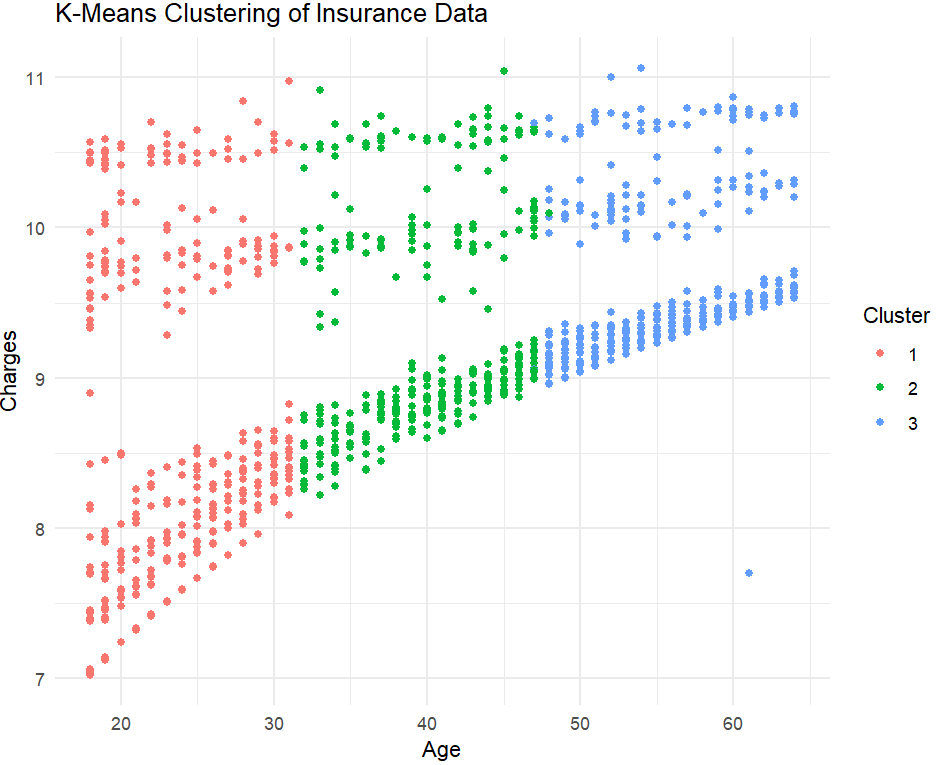
Elbow Point Indicates Optimal Clusters: The point where the reduction in WSS slows down and forms an "elbow" in the plot is a visual indicator of the optimal number of clusters. In this case, it occurs at three clusters.

Balance Between Complexity and Fit: Choosing three clusters strikes a balance between capturing meaningful patterns in the data and avoiding excessive complexity. Additional clusters may lead to overfitting and may not represent genuine structures in the data.

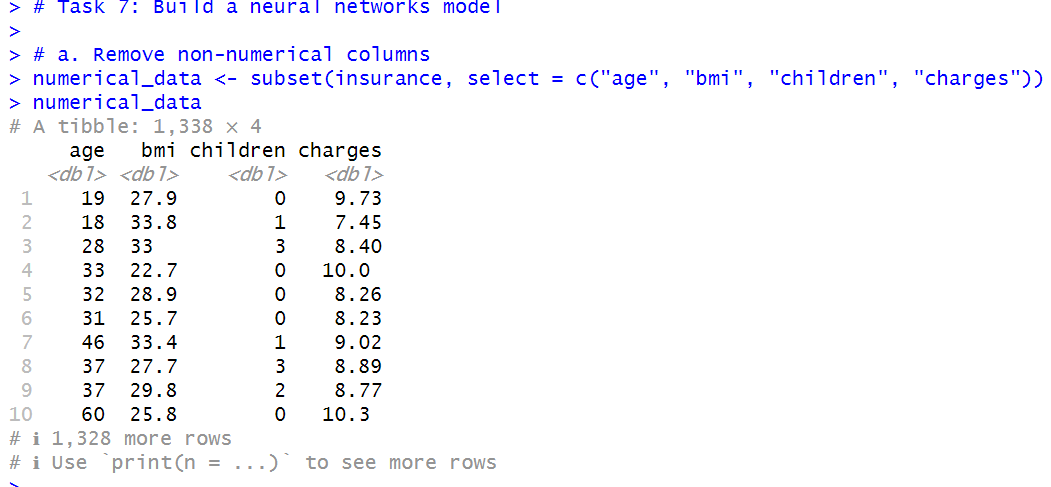
1. Perform k-means clustering using **3 clusters.**



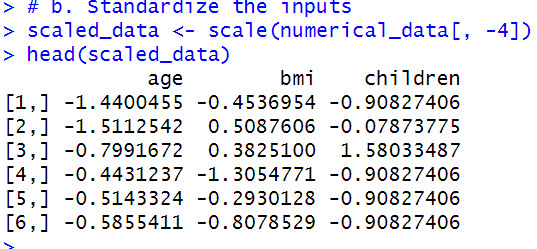
1. Visualize the clusters in different colors.



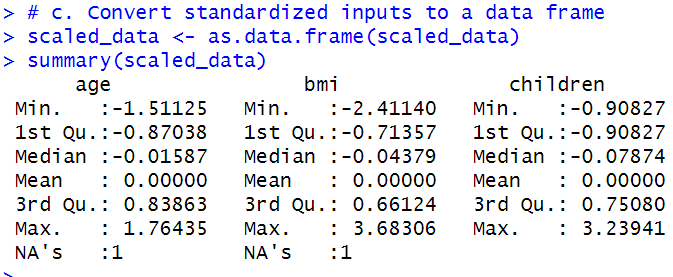
1. Build a neural networks model.
2. Remove the **sex, smoker,** and **region,** since they are not numerical values.



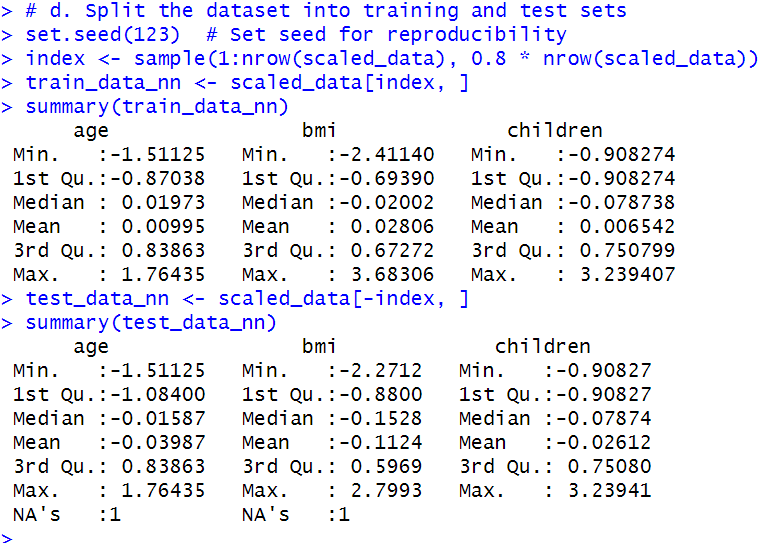
1. Standardize the inputs using the scale() function.



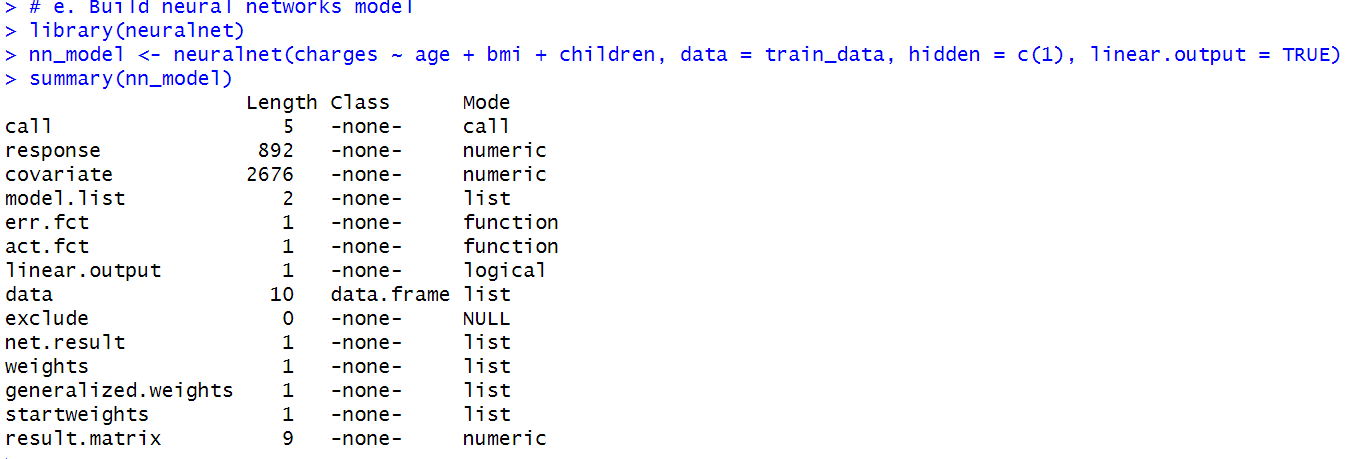
1. Convert the standardized inputs to a data frame using the as.data.frame() function.



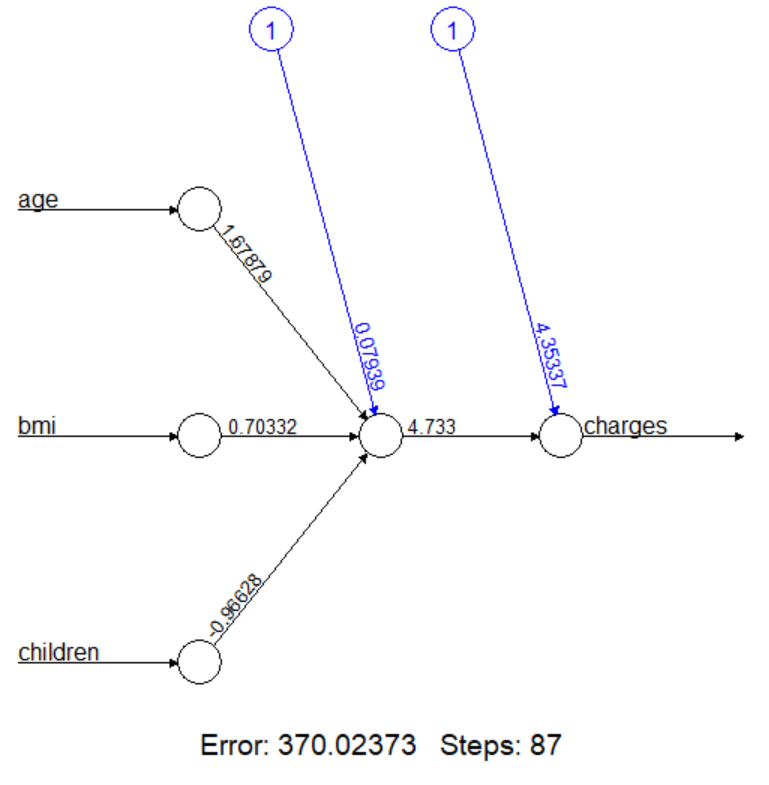
1. Split the dataset into a training set containing 80% of the original data and the test set containing the remaining 20%.



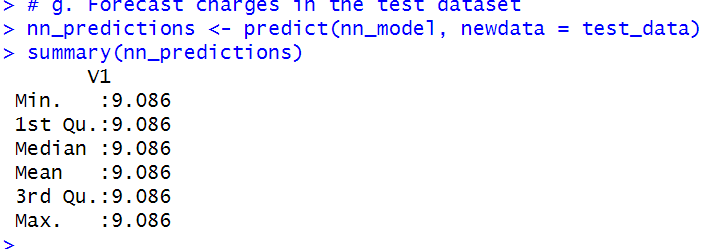
1. The response is **charges** and the predictors are **age, bmi,** and **children.**



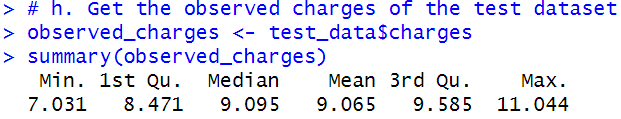
1. Plot the neural networks.



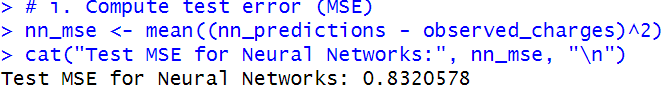
1. Forecast the charges in the test dataset.



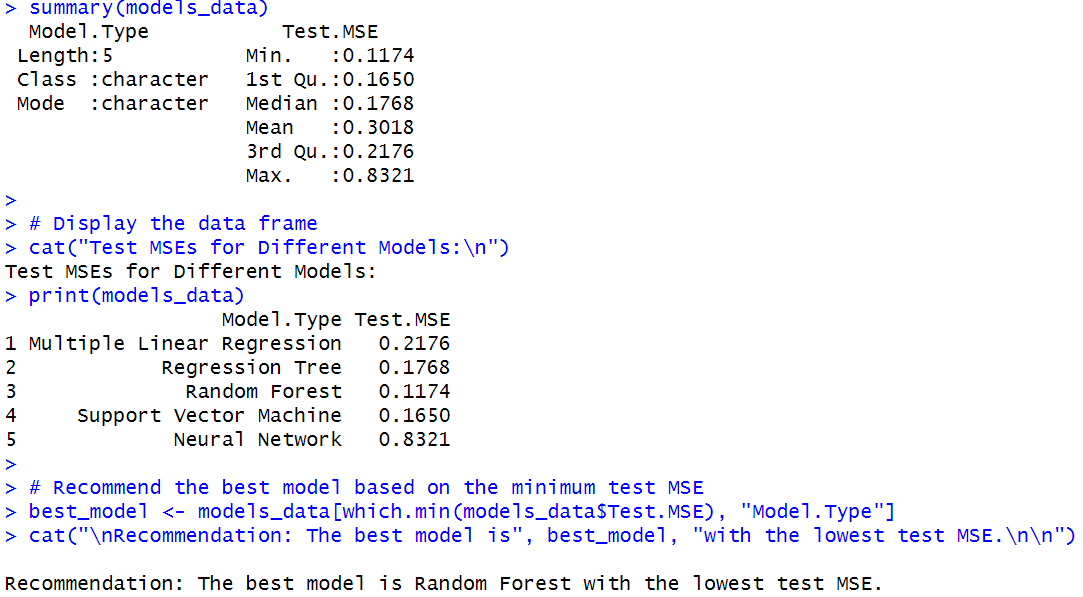
1. Get the observed charges of the test dataset.



1. Compute test error(MSE).



1. Putting it all together.
2. For predicting insurance charges, your supervisor asks you to choose the best model among the multiple regression, regression tree, random forest, support vector machine, and neural network models. Compare the test MSEs of the models generated in steps 2.g, 3.f, 4.b, 5.e, and 7.d. Display the names for these types of these models, using these labels: Multiple Linear Regression, Regression Tree, Random Forest, Support Vector Machine, and Neural Network and their corresponding test MSEs in a data.frame. Label the column in your data frame with the labels as Model.Type, and label the column with the test MSEs as Test.MSE and round the data in this column to 4 decimal places. Present the formatted data to your supervisor and recommend which model is best and why.



1. Another supervisor from the sales department has requested your help to create a predictive model that his sales representatives can use to explain to clients what the potential costs could be for different kinds of customers, and they need an easy and visual way of explaining it. What model would you recommend, and what are the benefits and disadvantages of your recommended model compared to other models?

Based on the test MSEs for different models, the Random Forest model performed the best among the models considered. Therefore, I would recommend the Random Forest model for predicting insurance charges. Here are some benefits and disadvantages of the Random Forest model:

Benefits:

1. High Predictive Accuracy: Random Forest models are known for their high predictive accuracy, often outperforming other models in a variety of scenarios.

2. Handles Non-Linearity: Random Forests can capture complex non-linear relationships in the data, making them suitable for datasets with intricate patterns.

3. Feature Importance: Random Forest provides a measure of feature importance, helping to identify the most influential variables in predicting the response.

4. Robust to Overfitting: The ensemble nature of Random Forest, combining multiple decision trees, helps reduce overfitting compared to a single decision tree.

Disadvantages:

1. Complexity and Interpretability: Random Forests consist of multiple trees, making them harder to interpret compared to simpler models like linear regression.

2. Computationally Intensive: Training a Random Forest can be computationally intensive, especially for large datasets and a large number of trees.

3. Hyperparameter Tuning: Random Forest has hyperparameters that need to be tuned for optimal performance. While they are less sensitive than some other models, tuning is still required.

4. Resource Intensive: In production, predicting with a Random Forest may require more resources compared to simpler models.

It's essential to consider the specific needs and constraints of the sales representatives when recommending a model. If interpretability is crucial, a simpler model like linear regression might be preferred, even if it sacrifices a bit of predictive accuracy. If accuracy is the top priority, Random Forest could be a suitable choice.

1. The supervisor from the sales department likes your regression tree model. But she says that the salespeople say the numbers in it are way too low and suggests that maybe the numbers on the leaf nodes predicting charges are log transformations of the actual charges. You realize that in step 1.b of this project that you had indeed transformed charges using the log function. And now you realize that you need to reverse the transformation in your final output. The solution you have is to reverse the log transformation of the variables in the regression tree model you created and redisplay the result.

